SUPPLEMENT

Software Application Profile

EpiMetal: An open-source graphical web browser tool for easy statistical analyses in epidemiology and metabolomics

Jussi Ekholm, Pauli Ohukainen, Antti J Kangas, Johannes Kettunen, Qin Wang, Mari Karsikas, Anmar A. Khan, Bronwyn A. Kingwell, Mika Kähönen, Terho Lehtimäki, Olli T. Raitakari, Marjo-Riitta Järvelin, Peter Meikle, and Mika Ala-Korpela

CONTENT:

Epidemiological data

Northern Finland Birth Cohort 1966

The Northern Finland Birth Cohort (NFBC66) was initiated in 1966 to study factors affecting preterm birth and subsequent morbidity in the two northernmost provinces in Finland (www.oulu.fi/nfbc). It included 12,058 children born into the cohort, comprising 96% of all births during 1966 in the region.[1-2] Data collection in 1997 included clinical examination and serum sampling at the age of 31 for 6,007 individuals. Serum NMR metabolomics data are available for 5,713 individuals. Data from this time point were used in the current analyses. Attendees in the 31-year field study (52%) were representative of the original cohort.[1] Informed written consent was obtained from all participants. The research protocols were approved by the Ethics Committee of Northern Ostrobotnia Hospital District, Finland.

Multiple funding bodies have supported the NFBC66 over the years, including European Regional Development Fund (Grant no. 539/2010 A31592) and the European Union's Horizon 2020 research and innovation programme (grant agreement no. 633595; DynaHEALTH).

Cardiovascular Risk in Young Finns Study

The Cardiovascular Risk in Young Finns Study (YFS) is an ongoing population-based follow-up study of atherosclerotic development.[3] The first cross-sectional survey was conducted in 1980 when 3,596 subjects aged 3–18 years participated in the study. Since then, follow-ups have been conducted in the whole population at regular intervals. In this study the 2001 (YFS01) and 2007 (YFS07) collections were used with the data, including the serum NMR metabolomics, available for 2,247 and 2,159 individuals, respectively. All participants in YFS provided written informed consent, and the study was approved by local ethics committees.

The YFS has been financially supported by the Academy of Finland: grants 286284, 134309 (Eye), 126925, 121584, 124282, 129378 (Salve), 117787 (Gendi), and 41071 (Skidi); the Social Insurance Institution of Finland; Competitive State Research Financing of the Expert Responsibility area of Kuopio, Tampere and Turku University Hospitals (grant X51001); Juho Vainio Foundation; Paavo Nurmi Foundation; Finnish Foundation for Cardiovascular Research ; Finnish Cultural Foundation; The Sigrid Juselius Foundation; Tampere Tuberculosis Foundation; Emil Aaltonen Foundation; Yrjö Jahnsson Foundation; Signe and Ane Gyllenberg Foundation; Diabetes Research Foundation of Finnish Diabetes Association; EU Horizon 2020 (grant 755320 for TAXINOMISIS); European Research Council (grant 742927 for MULTIEPIGEN project); and Tampere University Hospital Supporting Foundation.

Figures

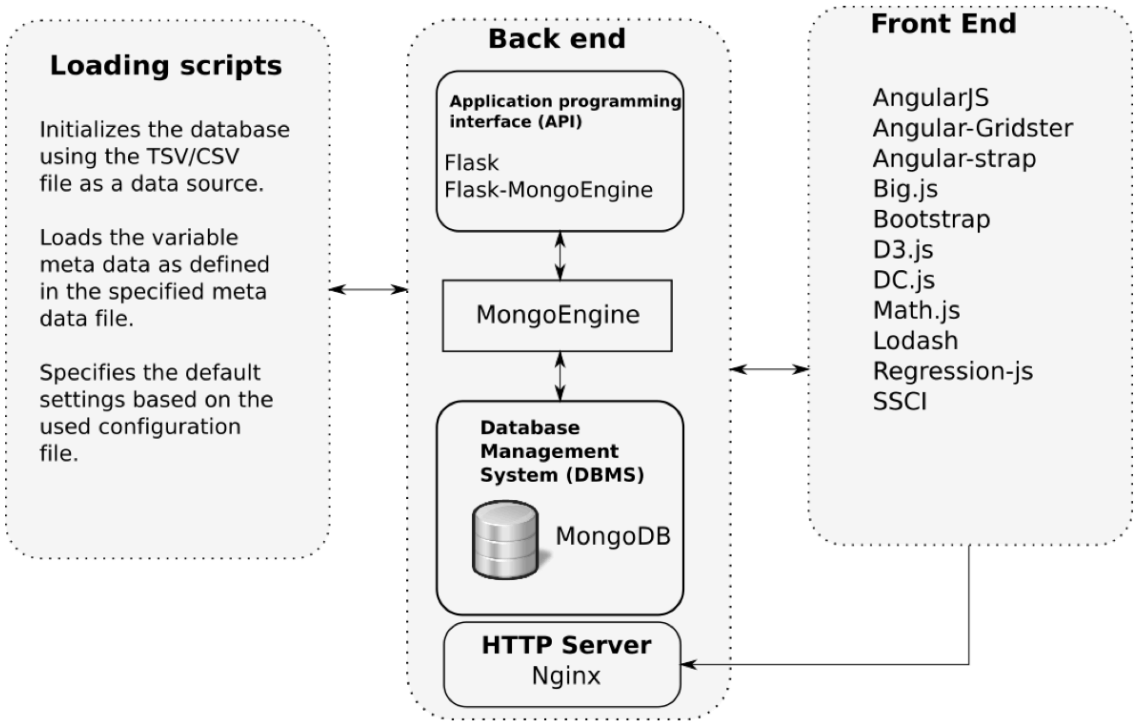


Figure S1. Overall software architecture of EpiMetal.

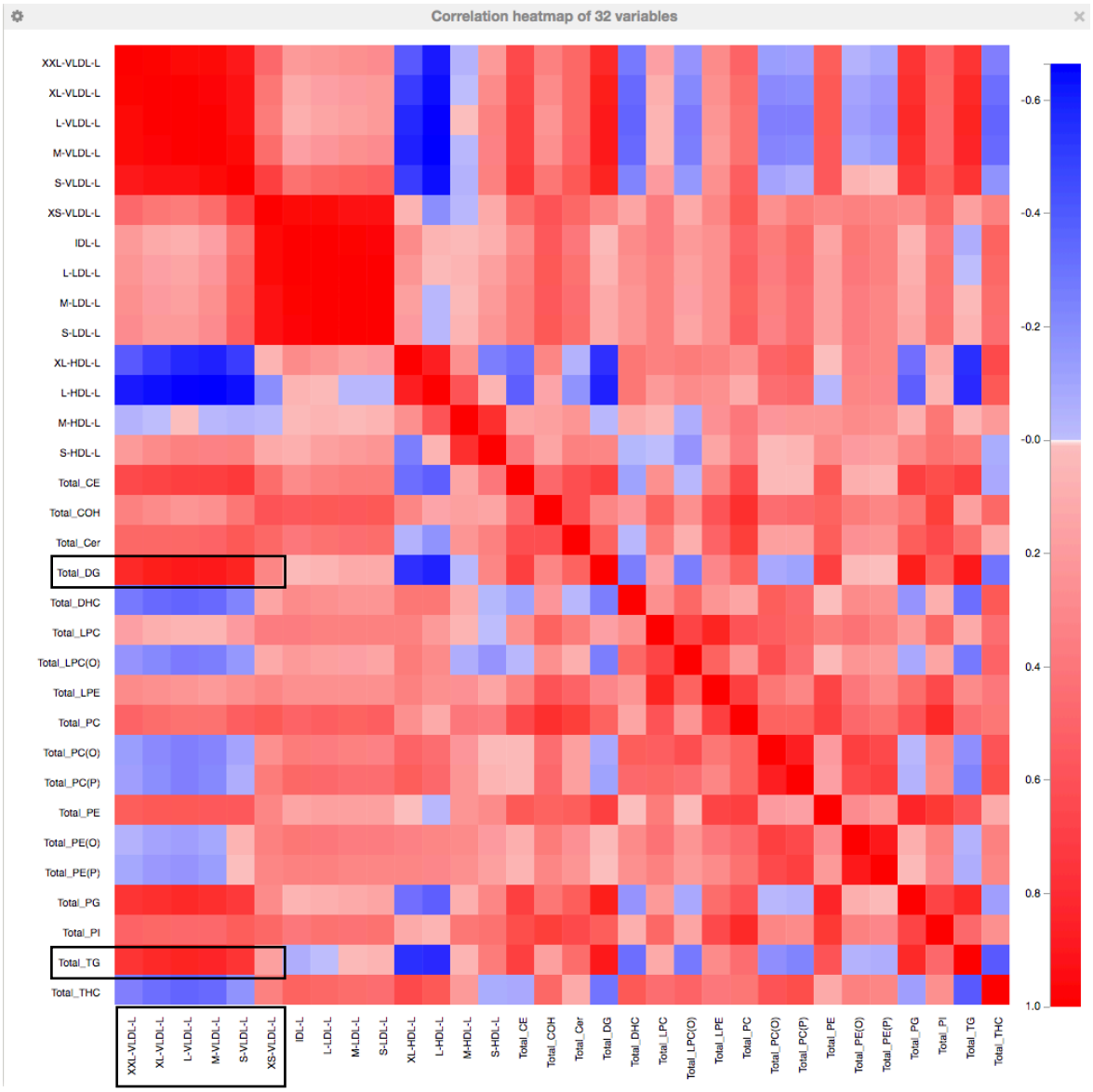


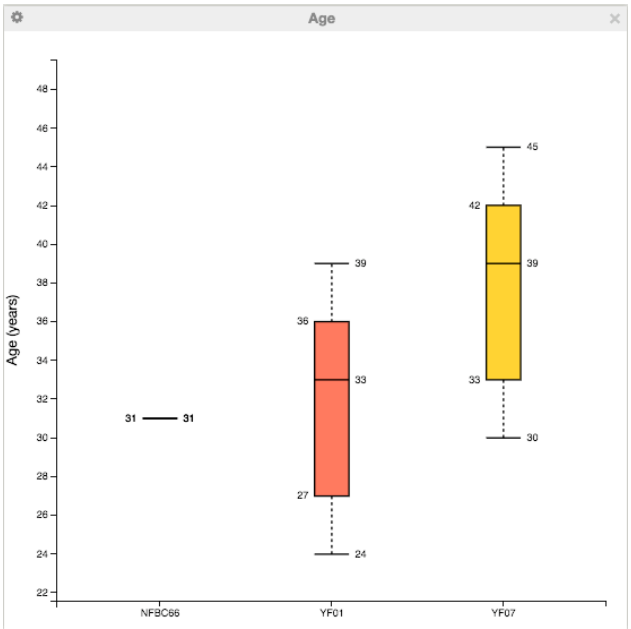
Figure S2. Correlation heatmap for lipoprotein subclasses (NMR spectroscopy) and individual lipid molecule classes (mass spectrometry lipidomics). These data are for 190 serum samples.[4-6] Both axes display the same selected measures from the NMR data[7,8] for the total lipids (L) in 14 lipoprotein subclasses (VLDL, very-low-density lipoproteins; IDL, low-density lipoproteins; LDL, low-density lipoproteins; HDL, high-density lipoproteins – XXL refers to the largest and XS to the smallest size) and from the mass spectrometry lipidomics[4] for the summary measures (Total) for the individual lipid molecule classes (CE, cholesterol esters; TG, triglycerides; COH, total cholesterol; Cer, ceramides; DG, diacylglycerols; DHC, dihexosylceramides; LPC, lysophosphatidylcholines; LPC(O), lysoalkylphosphatidylcholines; LPE, lysophosphatidylethanolamines; PC, phosphatidylcholines; PC(O), alkylphosphatidylcholines; PC(P), alkenylphosphatidylcholines; PE, phosphatidylethanolamines; PE(O),

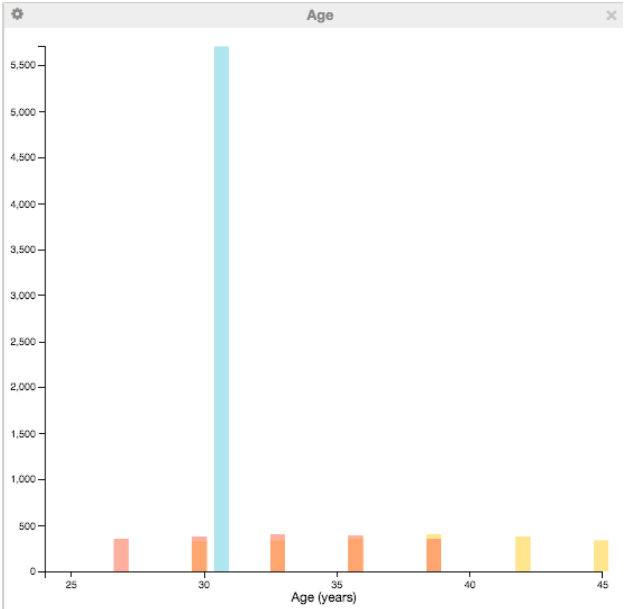
alkylphosphatidylethanolamines; PE(P), alkylphosphatidylethanolamines; PG, phosphatidylglycerols; PI, phosphatidylinositols; THC, trihexocylceramides). To simplify the presentation here, the individual lipid species (> 300 molecular measures) are not illustrated. The correlations are Spearman's rank correlations. The correlations are very well in line with the known molecular characteristics of lipoprotein subclasses and their compositions[9] and thereby demonstrate the robustness of the integrated data from these two independent metabolomics platforms. For example, highlighted in the figure are strong associations between diacylglycerols and triacylglycerols (mass spectrometry lipidomics) with total lipids in triglyceride-rich VLDL subclasses (NMR spectroscopy).

Illustration of EpiMetal usage in large population-based cohorts

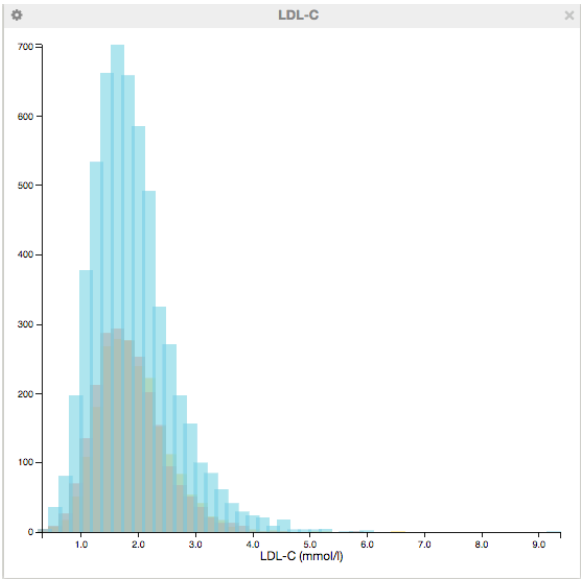
We also wanted to give a brief illustration of the general epidemiological suitability of the EpiMetal software. We uploaded the NFBC66 data (N=5,713, cyan) and data for two time points in the YFS (YFS01, N=2,247, orange and YFS07, N=2,159, yellow) into the EpiMetal software. Below some basic handling and analyses of these data are illustrated. It is of note that the cohort data can be analysed independently, combined into one (N=10,119) or any user-defined selection based on the self-organizing maps or other analyses.

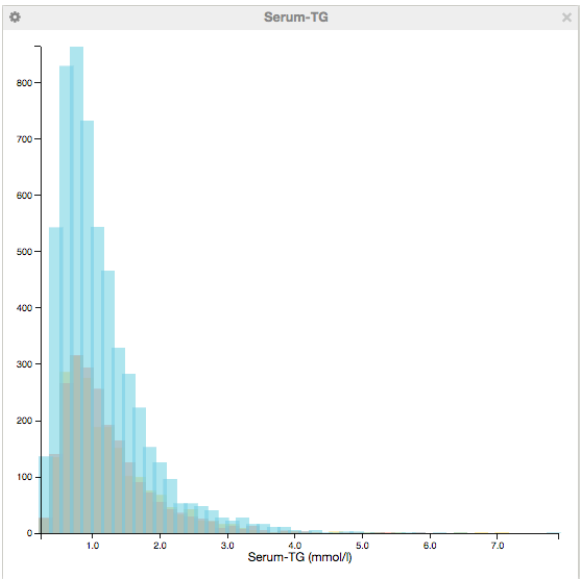
First, we plotted the age variable as a box plot and a histogram to highlight the different age distributions of the NFBC66 (a birth cohort with all the individuals within one year of age) and the YFS collections (with distinct age categories):



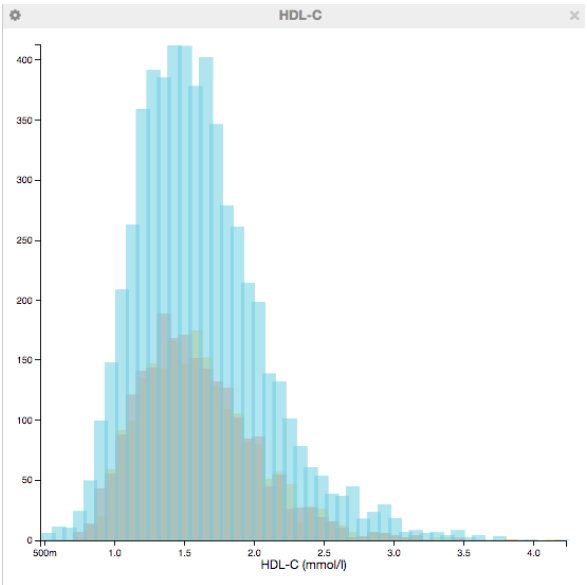


Then we plotted the distributions of key lipoprotein lipid concentrations, low-density lipoprotein cholesterol (LDL-C), triglycerides (TG) and high-density lipoprotein cholesterol (HDL-C) in each data set on top of each other:

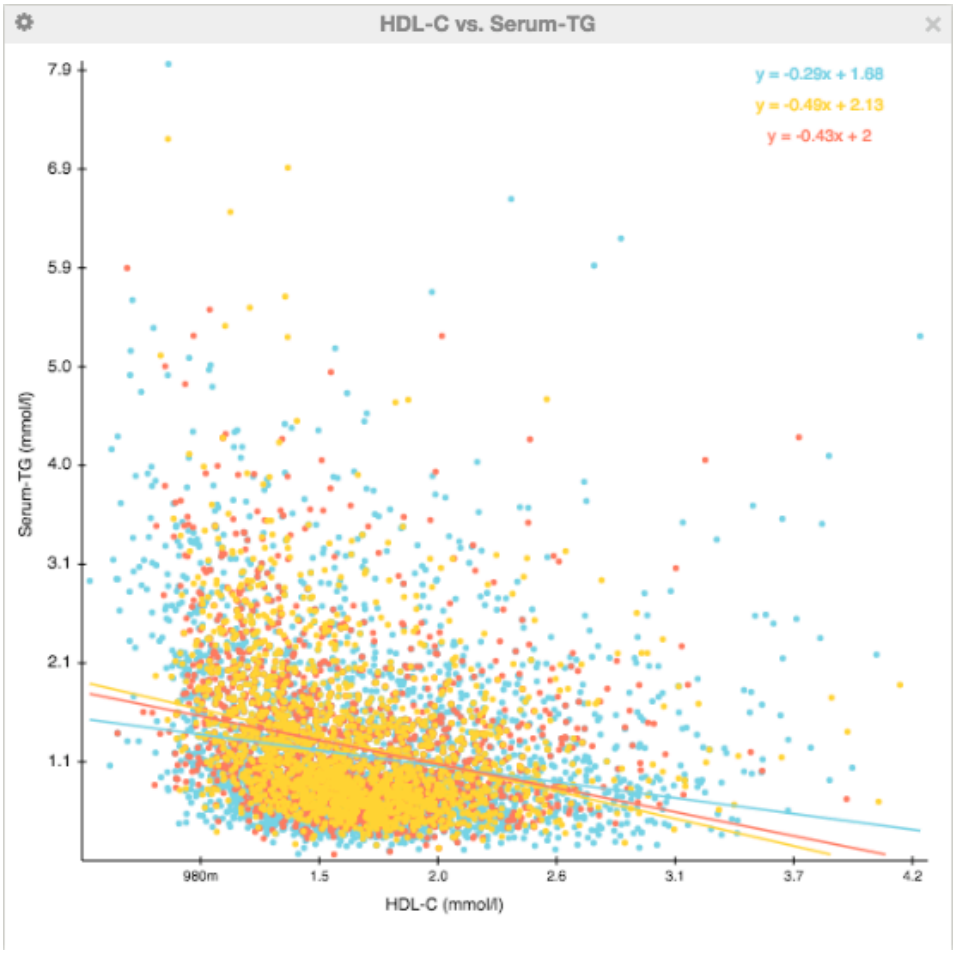




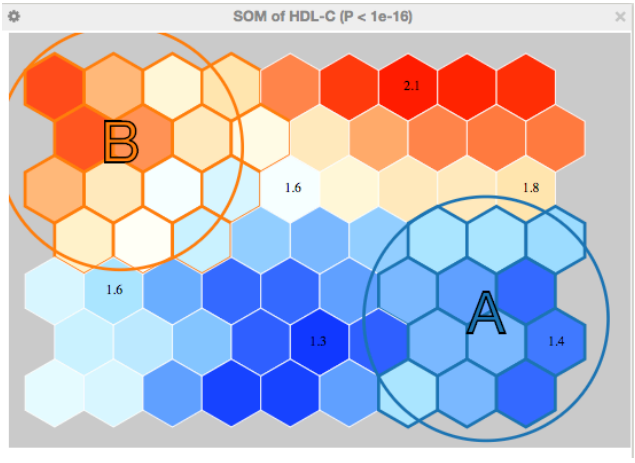
The distributions illustrate the expected closely normal distributions for LDL-C and HDL-C in large population samples. In addition, as expected, the distributions for TG are positively skewed in all data sets. The distributions are also heavily overlapping due to the homogeneity of the group (all 'young Finns').

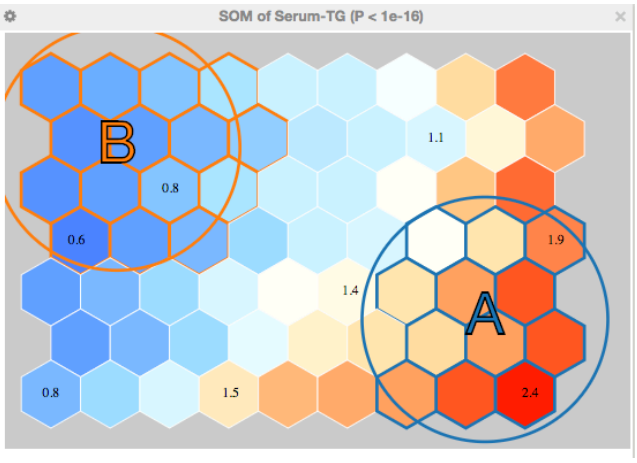


Next, we replicated the expected negative population-level negative correlation between circulating serum TG and HDL-C; again, all data behave coherently:

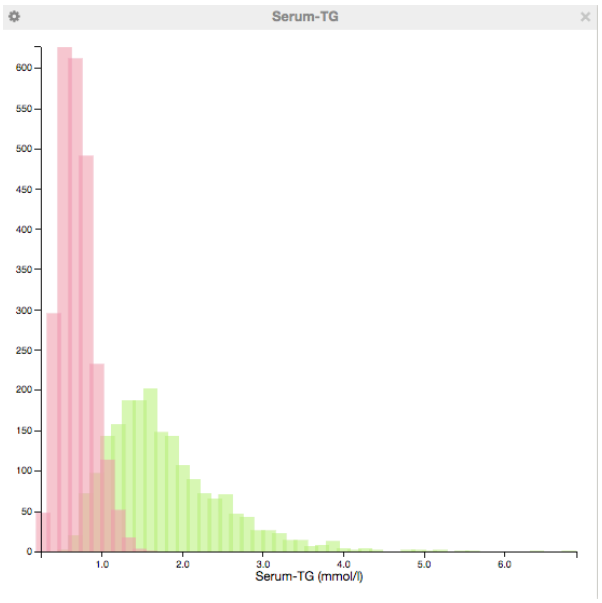


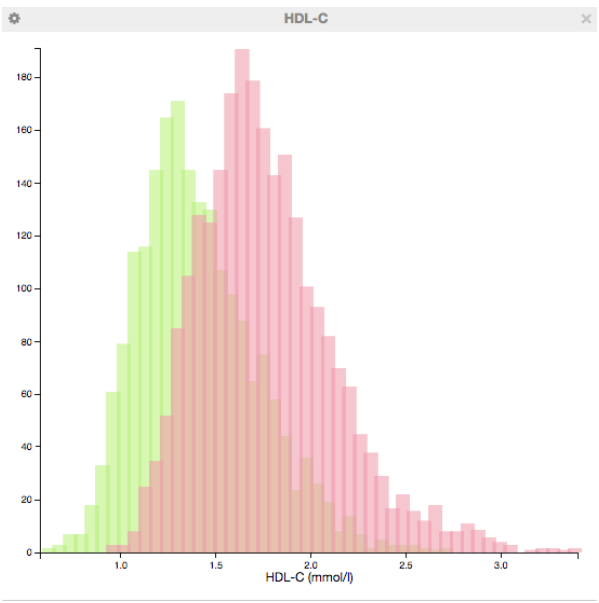
Then we recalculated a combined SOM of all the data (N=10,119): individuals with low concentrations of TG typically have high levels of HDL-C (circle B) and vice versa (circle A), recapitulating the above scatterplot and correlation:





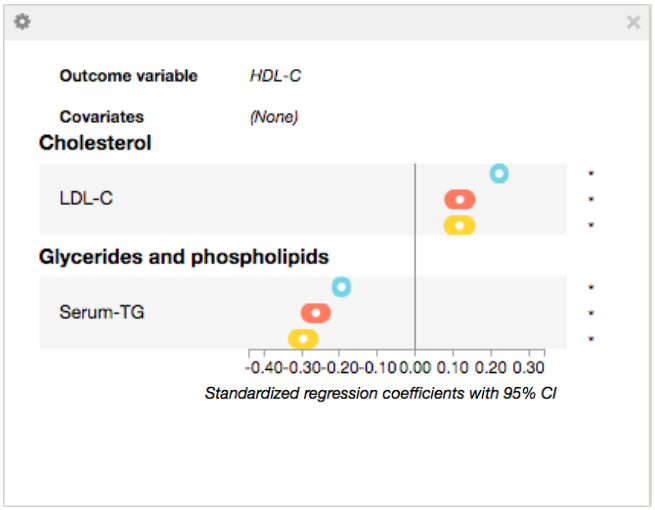
It is a key point that the selections within the SOM represent distributions as demonstrated by the following plots for the circulating lipid concentrations for the individuals in the selected areas (green for A and pink for B):

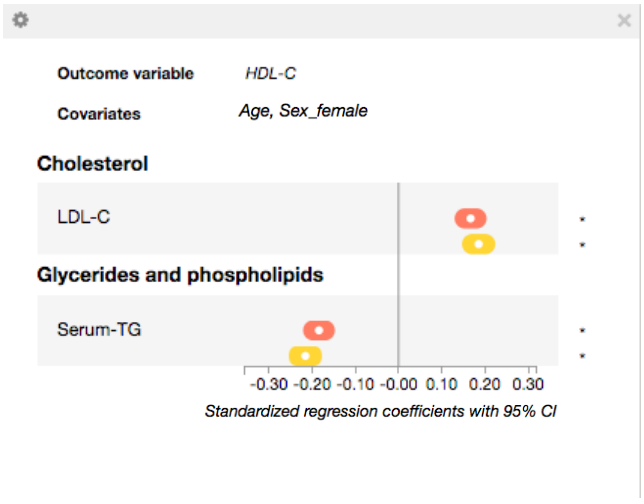




The TG and HDL-C distributions based on the SOM categories remain overlapping but have clearly separate mean values as expected by the SOM-based phenotypes.

Looking at the data via formal regression analyses reveal the same issues, of course:





These plots also emphasise the utility for inclusion of covariates in regression analyses using EpiMetal; in this example for the NFBC66 the software automatically excludes the data when trying to adjust for age (that is basically a constant for a birth cohort). This omission is prompted to the user by the EpiMetal software.

References

1. Järvelin M, Sovio U, King V, *et al*. Early life factors and blood pressure at age 31 years in the 1966 northern Finland birth cohort. *Hypertension* 2004;**44**: 838-46.
2. Sabatti C, Service SK, Hartikainen AL, *et al*. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat Genet* 2009;**41**:35-46.
3. Raitakari OT, Juonala M, Rönnemaa T, *et al*. Cohort profile: the cardiovascular risk in Young Finns Study. *Int J Epidemiol* 2008;**37**:1220–6.
4. Khan AA, Mundra PA, Straznicky NE, *et al*. Weight Loss and Exercise Alter the High-Density Lipoprotein Lipidome and Improve High-Density Lipoprotein Functionality in Metabolic Syndrome. *Arterioscler Thromb Vasc Biol* 2018;**38**:438–47.
5. Straznicky NE, Lambert EA, Nestel PJ, *et al*. Sympathetic Neural Adaptation to Hypocaloric Diet With or Without Exercise Training in Obese Metabolic Syndrome Subjects. *Diabetes* 2010;**59**:71–9.
6. Straznicky NE, Grima MT, Sari CI, *et al*. A Randomized Controlled Trial of the Effects of Pioglitazone Treatment on Sympathetic Nervous System Activity and Cardiovascular Function in Obese Subjects With Metabolic Syndrome. *J Clin Endocrinol Metab* 2014;**99**:E1701–7.
7. Soininen P, Kangas AJ, Würtz P, Suna T, Ala-Korpela M. Quantitative serum nuclear magnetic resonance metabolomics in cardiovascular epidemiology and genetics. *Circ Cardiovasc Genet* 2015;**8**:192–206.
8. Würtz P, Kangas AJ, Soininen P, Lawlor DA, Davey Smith G, Ala-Korpela M. Quantitative Serum Nuclear Magnetic Resonance Metabolomics in Large-Scale Epidemiology: A Primer on -Omic Technologies. *Am J Epidemiol* 2017;**186**:1084–96.
9. Kumpula LS, Kumpula JM, Taskinen MR, Jauhiainen M, Kaski K, Ala-Korpela M. Reconsideration of hydrophobic lipid distributions in lipoprotein particles. *Chem Phys Lipids* 2008;**155**:57-62.